COMP571 (Fall 2020)

Important note: The information contained in the course syllabus, other than the absence policies, may be subject to change with reasonable advance notice, as deemed appropriate by the instructor.

Who

Instructor:

- Huw A. Ogilvie
- Duncan Hall 3098
- hao3@rice.edu

TAs:

- TBA

Where and when

This year, in-person attendance will be entirely optional. Your choice to take COMP571 online should make as little difference to your experience or results as possible. You can change your mind during the semester and stop or start in-person attendance. The situation is dynamic and in-person attendance may or may not be possible for the entire semester, but online attendance will always remain an option.

Distribution of class materials and submission of assignments and projects will be conducted via Canvas.

Lectures will be held in Duncan Hall **1046**, on Tuesdays and Thursdays, between 3:10–4:30 PM. If you want to attend lectures in-person, you will have to nominate whether you want to attend on Tuesday **or** Thursdays. To ensure reduced class sizes for social distancing, you should **not** attend both days in-person. All lectures will be recorded and available online.

One scheduled office hour will be held each week on Zoom, on a day and time to be announced, and attendance by the whole class is encouraged so that everyone can benefit from the discussion. Office hours will not be recorded. Individual appointments outside this time are welcome.

Intended audience

The students who should take COMP571 are generally studying computer science, biology or genomics, and wish to learn how to apply algorithms and statistical models to important problems in biology and genomics.

Course objectives and learning outcomes

The primary objective of the course is to teach the theory behind methods in biological sequence analysis, including sequence alignment, sequence motifs, and phylogenetic tree reconstruction. By the end of the course, students are expected to understand and be able to write basic implementations of the algorithms which power those methods.

Course materials

The main material for this course will be lectures and the course blog. The text for Professor Treangen's course *Genome-Scale Algorithms* is *Bioinformatics Algorithms* by Compeau & Pevzner, which contains relevant chapters and is now available for free online. However the focus of COMP571 is on the nexus of sequence analysis and statistical models, whereas the focus of Bioinformatics Algorithms and Professor Treangen's course is on algorithms and data structures.

Software for the course

Algorithms and statistics will be demonstrated using Python. Assignments and projects will require some Python coding. R may be used for some demonstrations (because it is nice for data visualization) but not for assessment.

The NumPy and SciPy libraries for scientific computing will be used with Python. To install these libraries, first install the latest official distribution of Python 3. This can be downloaded for macOS or for Windows from Python.org, and should already be included with your operating system if you are using Linux.

Then simply use the Python package manage pip to install NumPy and SciPy from the command line, by running `pip3 install numpy scipy`.

Schedule and assessment

The course is organized around three themes, and there will be a corresponding homework assignment for each one;

1. Models and algorithms used for sequence alignment
2. Hidden Markov Models in computational biology
3. Phylogenetic and coalescent inference

In addition to these assignments, each student will have to complete one project of implementing a novel or existing statistical model, applying it to a public data set, and writing up the results in the style of a scientific paper. The statistical model should be relevant to one (or more) of the course themes. Projects will be designed by groups of 5-10 students, but the implementation, application and write up will be individual. The due date of the project is determined by the theme the group chooses to focus on.

Project design and discussion will take place either on Zoom or in socially distanced outdoor environments depending on the preference and physical location of students in each group.

The below schedule may change subject to Rice University policy

Monday's date	Tuesday's lecture	Thursday's lecture	Homework	Project
08/24/20	Introduction	Gene models and central dogma		
08/31/20	PSSMs (gapless motifs)[1]	Pseudocounts[1]		
09/07/20	BLOSUM and PAM[1]	Likelihood ratios[1]		
09/14/20	Global alignment[1]	Local alignment and BLAST[1]	#1 issued	
09/21/20	Sequence alignment review[1]	Hidden Markov models[2]		
09/28/20	Viterbi algorithm[2]	Gapped motifs[2]	#1 due	
10/05/20	Forward algorithm[2]	Backward algorithm[2]		
10/12/20	HMM review[2]	Phylogenetic trees[3]	#2 issued	#1 due
10/19/20	Equal-cost parsimony[3]	Unequal-cost parsimony[3]		
10/26/20	Hill climbing[3]	The Felsenstein zone[3]	#2 due	
11/02/20	Likelihood of two sequences[3]	Felsenstein's pruning algorithm[3]		
11/09/20	The Wright-Fisher model[3]	Coalescent theory[3]	#3 issued	#2 due
11/16/20	Markov chain Monte Carlo[3]	Phylogenetics review[3]		
11/23/20	*No instruction*	*No instruction*	#3 due	
11/30/20	*No instruction*	*No instruction*		
12/07/20	*No instruction*	*No instruction*		#3 due

Superscript numbers refer to the theme(s) for that day's class or midterm. Assignments will be issued before midnight on Sundays. Assignments and projects will also be due before midnight on Sundays.

Grade policies

- Homework assignments: 20% each
- Project design: 10%
- Project implementation: 10%
- Project report: 20%

Assignments or projects submitted late with a strong and valid excuse will be accepted without penalty. The strength and validity of excuses will be solely the instructor's purview. Without a strong and valid excuse, the final course percentage will be reduced by 2% for each day any submission is late, up to the contribution of that submission to the final percentage. For example if submitted homework is given a mark of 70%, it contributes 70% × 20% = 14% to the final percentage.

No assignments or projects will be accepted after the end of the semester on Wednesday, December 16, 2020. In exceptional circumstances, if a student is unable to complete an assignment or project before the semester ends, the final percentage will be calculated by scaling the assessment which that student has completed. Again, this will be solely within the instructor's purview.

Absence policies

Please stay safe and healthy. Do your best to either attend or view lectures and participate in project meetings.

Rice Honor Code

In this course, all students will be held to the standards of the Rice Honor Code, a code that you pledged to honor when you matriculated at this institution. If you are unfamiliar with the details of this code and how it is administered, you should consult the Honor System Handbook at http://honor.rice.edu/honor-system-handbook/. This handbook outlines the University's expectations for the integrity of your academic work, the procedures for resolving alleged violations of those expectations, and the rights and responsibilities of students and faculty members throughout the process.

Students with a disability

If you have a documented disability or other condition that may affect academic performance you should: 1) make sure this documentation is on file with Disability Support Services (Allen Center, Room 111 / adarice@rice.edu / x5841) to determine the accommodations you need; and 2) talk with me to discuss your accommodation needs.